Modeling correlations among air pollution-related data through generalized association rules

Modeling correlations among air pollution-related data through generalized association rules

Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, Giuseppe Ricupero, Xin Xiao

Dipartimento di Automatica e Informatica,
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
E-mail: *luca.cagliero@polito.it*

Keywords Smart Cities, Data Mining, Pollutant Data, Sensor Networks.

Abstract *Today's citizens and city administrations have an increasing interest in monitoring the air quality in urban areas. Studying the causes of air pollution entails analyzing the correlations between heterogeneous data, among which pollutant concentrations, traffic flow measurements, and meteorological data. To this end, innovative data analytics solutions able to acquire, integrate, and analyze very large amounts of data are needed. This paper presents a new data mining system, named* GE*neralized* C*orrelation analyzer of p*O*llution data (*GECKO*), to discover interesting and multiple-level correlations among a large variety of open air pollution-related data. Specifically, correlations among pollutant levels and traffic and climate conditions are discovered and analyzed at different abstraction levels. The knowledge extraction process is driven by a taxonomy to generalize low-level measurement values as the corresponding categories. To ease the manual inspection of the result, the extracted correlations are classified into few classes based on the semantics of underlying data. The experiments, performed on real data acquired in a major Italian Smart City, demonstrate the effectiveness of the proposed analytics engine in discovering correlations among pollutant data that are potentially useful for supporting city administrators in decision-making.*

1 Introduction

In the last few years various sensor networks have been deployed in smart cities to collect a variety of data of public interest. Such data have a great potential to influence the quality of life of urban city dwellers. Collected data can be analyzed to discover knowledge useful for driving public administrators in decision-making and, thus, enhancing the quality of life of citizens. This paper addresses a key issue in smart city environments, i.e., the monitoring and analysis of the air quality in urban areas.

Air pollution may have a serious impact on the public health. The quality of the air can vary over time and across different areas of the same city. Furthermore, it is influenced by different factors such as the weather conditions (e.g. humidity, temperature and atmospheric pressure) and human activities (e.g., traffic flows, people's mobility). To monitor pollutant concentrations and their relationship with meteorological and traffic conditions, sensor networks are deployed by the public administration over the city area. Air quality data acquired from network sensors can be further enriched with that acquired by recently wearable sensors, while climate data can be measured through personal weather stations.

The evaluation of how the above factors impact on the air quality is currently a relevant research issue. Previous works have already studied the correlation between different pollutants through statistics-based methods such as one-way ANOVA analysis Lodovici et al. [2003]. Furthermore, Principal Component and Canonical Correlation analyses Statheropoulos et al. [1998] have been exploited to analyze the correlation between pollutants and meteorological data Elminir [2005]. A parallel effort has been devoted to exploiting data mining techniques to analyze the air quality levels in urban environments Zheng et al. [2013, 2015]. Classification algorithms have been exploited to predict the air quality level in areas not equipped with monitoring stations Zheng et al. [2013]. To train the classification model, historic and real-time measurements on air quality, weather conditions, traffic flows, and people's mobility have jointly been analyzed. Similarly, in Zheng et al. [2015] air quality and meteorological data acquired in the past were analyzed to predict the level of the air quality in the near future.

Association rule mining approaches have found application in various application domains (e.g. network traffic analysis Baralis et al. [2010], social data analysis Cagliero and Garza [2013]) to discover interesting correlations among data items. The exploitation of these approaches on air pollution-related data can support the discovery of interesting yet hidden knowledge. The extracted patterns are commonly managed by domain experts through manual inspection to support decision-making.

This paper presents a data mining system, named GEneralized Correlation analyzer of pOllution data (GECkO), to extract interpretable correlations, at different abstraction levels, among a large variety of data related to air quality. Pollutant measurements are first integrated with traffic and meteorological data and enriched with an analyst-provided taxonomy, which aggregates measurement values into the corresponding higher-level categories. Then, an established generalized association rule mining algorithm Baralis et al. [2010] is applied to the prepared dataset. The extracted rules, namely the generalized association rules, represent frequent co-occurrences between pollutant levels and environmental conditions at different abstraction levels. Finally, to ease the expert-driven rule inspection process, the rules are classified into few classes according to the semantics of the represented information.

The GECkO system was validated on real data collected in a major Italian city. The discovered patterns demonstrate the effectiveness of GECkO in discovering interesting knowledge that can be easily exploited by public administrators to monitor the air quality in urban environments.

This paper is organized as follows. Section 2 thoroughly describes the GECkO system. Section 3 summarizes the performed experiments, while Section 4 draws conclusions and discusses future research directions.

2 The GEneralized Correlation analyzer of pOllution data architecture

The GEneralized Correlation analyzer of pOllution data (GECkO) system is a data mining engine that analyze the correlations between pollutants and different environmental factors, such as meteorological and traffic conditions, in a Smart City context. The main architectural blocks are:
(i) *Data integration*, in which pollutant and environmental data are acquired and integrated,
(ii) *Data representation*, in which data are tailored to a relational data format and enriched with a taxonomy aggregating concepts into higher-level ones,
(iii) *Data analyses*, in which generalized association rules are extracted from the prepared data to support domain experts in performing advanced analyses.

A more detailed description of each block is given in the following.

2.1 Data integration

Since the concentrations of pollutants can be relevantly affected by both weather conditions (e.g., temperature, humidity) and type of traffic crossing the city area (e.g. how many gasoline engine vehicles crossed the area), different sensor networks should be exploited to periodically monitor values for different data types. Specifically, measurements for three main types of data should be acquired: pollutant data, meteorological data, and traffic data. In urban environments, a different geo-referenced sensor network is usually deployed for monitoring each of the above data types. An ad hoc integration strategy is applied since the considered sensor networks may adopt a different timeline in sampling values and be deployed in different city areas. In the following we first describe the considered data types, and then the data integration strategy currently adopted in GEneralized Correlation analyzer of pOllution data.

Pollutant data. Concentration measurements for each pollutant were periodically collected through dedicated sensors deployed in pollution monitoring stations (PolMS). Each station is characterized by the geo-coordinates (i.e., latitude and longitude) of its location, and stations are located in different areas of the city. The most damaging pollutants are monitored, including particulate matters PM_{10} and $PM_{2.5}$, carbon monoxide (CO), and ozone (O_3). Each station monitors the concentrations of various pollutants at a fixed time granularity. Depending on the type of pollutant, the frequencies of data acquisition can be hourly or daily.

Meteorological data. To analyze the climate conditions of the urban area, the GEneralized Correlation analyzer of pOllution data collects the most common meteorological indicators (e.g. air temperature, relative humidity, precipitation level, wind speed, atmospheric pressure). Climate conditions are acquired through geo-referenced meteorological stations distributed throughout the urban territory.

Traffic data. The concentration of traffic is measured as the number of vehicles entering a city area at a given time granularity (e.g. hourly). Since vehicles equipped with different engines may affect the air quality differently, we considered traffic data separately for each category of vehicles. Specifically, vehicles are categorized based on their fuel type (e.g., gasoline, diesel, electric).

To allow the analysis of the correlations between pollutant levels and environmental factors (i.e., weather and traffic conditions), the three different types of data described above are integrated into a unique repository. Meteorological and traffic data are preprocessed before data integration to align the spatial and temporal granularity of the acquired data. Since the analysis is focused on pollutant data, the spatial-temporal granularity of the sensor network monitoring pollutant concentrations is considered as a reference for time and space alignment.

To effectively deal with alignment issues, for each Pollution Monitoring Station (PolMS) meteorological and traffic data are aligned to the closest timestamp available in pollutant data through an approximate join. Specifically, meteorological data associated with a given pollution station are computed as a distance-based weighted mean of the values provided by the three nearest meteorological stations monitoring climate data. The weight assigned to each value is inversely proportional to the distance from these three stations to the PolMS. Hence, three equally distant meteorological stations would have the same importance for determining the weather values of a given city area. For traffic data the number of vehicles entering each area is associated to all the sensors deployed in the area. Traffic data are timely integrated through an approximate join similar to that adopted for climate data integration.

2.2 Data representation

To perform association rule-based analyses, heterogeneous data acquired from sensors are tailored to a relational data format, prepared to the next mining step by means of established preprocessing techniques, and enriched with a taxonomy, which generalized the relational model to a multiple-level model.

Relational data model A relational dataset is a set of records. Each record r_i corresponds to a given time period T_i and it collects pollutant, meteorological, and traffic data acquired in T_i. A record is a set of items, where an item is a pair (*attribute*, *value*). While *attribute* is the description of a data feature of interest in the context under analysis, *value* is the value assumed by the corresponding attribute. Each record contains at most one item per data attribute (i.e., multiple attribute values in the same record are not allowed).

In our context of analysis, we will consider the following attributes. (i) Pollutants: particulate matters PM_{10} and $PM_{2.5}$, Ozone (O_3), Nitrogen dioxide (NO_2), Carbon Monoxide (CO), and Benzene C_6H_6. (ii) Meteorological factors: wind direction, wind speed, temperature, humidity, pressure, UV radiations, precipitations. (iii) Traffic conditions: numbers of gasoline engine, diesel engine, natural gas, electric, and hybrid vehicles.

Data discretization Continuous attributes are unsuitable for use in association rule-based analyses, because their values are very unlikely to frequently occur in the analyzed dataset. For this reason, a data discretization step is applied prior to running the association rule mining process.

Pollutant concentration levels are discretized into different categories named with colors from green to red according to the severity of the level range from the point of view of the citizen's health. Currently, categories have been defined based on the classification given the Italian ARPA Piemonte agency responsible for environment protection in the Piemonte region ARPA. Piedmont Region (e.g. *blue* and *green* imply non-critical levels, while *orange* and *red* indicate highly critical levels).

The *traffic indicator* values are uniformly discretized by using the equal-width discretization algorithm available in the RapidMiner suite Rap [2016]. For example, the humidity values (expressed in $\frac{kg}{m^3}$) are discretized as *very low* between zero and 20, *low* between 20 and 40, *medium* between 40 and 60, *high* between 60 and 80, *very high* between 80 and 100, while for the UV radiations (expressed in $\frac{W}{m^2}$) the discretization levels are the following ones: *very low* between zero and 0.9, *low* between 0.9 and 2.9, *medium* between 2.9 and 5.9, *high* between 5.9 and 7.9, *very high* between 7.9 and 10.9, *extremely high* above 10.9.

Concerning the *meteorological attributes*, the wind speed is discretized, according to the Beaufort scale, in 13 different levels, from *Calm* (level 0) to *Hurricane force* (level 12), while the other attributes are discretized into standard value ranges. For example, the wind direction degrees are discretized based on the classical cardinal points (i.e., as *north-east*, *east*, *south-east*, *south*, *south-west*, *west*, *north-west*, and *north*).

Taxonomy generation To analyze pollutant data at different abstraction levels a taxonomy is built on top of relational data. A taxonomy is a set of is-a hierarchies, each one referring to a specific data attribute. Each hierarchy aggregates all the values assumed by the corresponding attributes into higher-level concepts in a tree-based structure. For example, let us consider the wind direction attribute. Low-level (discrete) values *north-east*, *east*, and *south-east* are generalized as *east-side*, while values *south-west*, *west*, *north-west* are generalized as *west-side*. An item consisting of a pair (*attribute*, *generalized value*), where *generalized value* is an higher-level aggregation occurring in the input taxonomy, will be hereafter denoted as *generalized item*. For example, based on the hierarchy on the wind direction attribute, item (*wind direction*, *north-west*) can be generalized as the corresponding generalized (higher-level) item (*wind direction*, *west-side*).

Taxonomies are analyst-provided. They can be either given by the domain expert based on their common knowledge or generated semi-automatically by applying multiple discretization runs on the same attribute domain. To generate the taxonomy, further discretization runs on top of discretized record values are applied. Pollutant concentration level categories (e.g., *blue* and *green*) are further discretized as *non-critical*, *fairly-critical*, and *highly critical* according to the level of severity of the pollutant from the point of view of the citizen's health. Traffic levels are discretized as *low*, *medium*, and *high*. Meteorological values are further discretized into upper-level categories (e.g. *east-side*, *west-side*). Hourly timeslots are categorized as 4-hour, and 8-hour timeslots (e.g., early morning, evening), while dates are aggregated into the corresponding week of the month (e.g. 1st week of December) , month of the year (e.g., December), and season (e.g., winter).

Since the process of taxonomy generation is semi-automatic, the taxonomy may consist of hierarchies of different height. To avoid bias in the next association rule mining process, the hierarchies in the taxonomy are balanced by equalizing the corresponding heights. As discussed in Cagliero et al. [2014], the aforementioned procedure is established in generalized pattern mining. To this aim, artificial root nodes are added to lower-height hierarchies until all their heights match those of the highest one.

2.3 Data analyses

This block aims at discovering interesting associations between pollutant levels and environmental factors (meteorological and traffic conditions), in the form of generalized association rules. Association rule mining Agrawal and Srikant [1994] is an exploratory data mining technique that has largely been used to extract hidden correlations among data items from large datasets.

Preliminary definitions To introduce the concept of association rule, we first recall the notion of itemset. In the context of relational data, an *itemset* is a set of items (*attribute*, *value*) all belonging to distinct attributes. For example, itemset {($PM_{2.5}$, *red*),(*wind-direction*, *south-east*)} indicates that items ($PM_{2.5}$, *red*) and (*wind-direction*, *south-east*) co-occur in the analyzed data.

To analyze pollutant data at different granularity levels, the itemset definition can be straightforwardly extended to the case in which data are enriched with a taxonomy. A generalized itemset Srikant and Agrawal [1995] is defined as a set of items and/or generalized items. Note that traditional (non-generalized) itemsets are special case of generalized itemset in which all items assume non-aggregated values according to the input taxonomy. For example, generalized itemset {($PM_{2.5}$, *highly critical*),(*wind direction*, *east-side*)} generalizes the former itemset by aggregating item values according to the hierarchies built on the $PM_{2.5}$ and *wind-direction* attributes (see Section 2.2).

A generalized item *matches* a given record if its value corresponds or is an aggregation of the value of any item of the record (at any abstraction level). For example, generalized item (*date*, *Winter*) matches a record

containing item (*date*, *December 1st, 2013*). The support of a generalized itemset in a relational dataset is an established quality index which is computed as the percentage of dataset records matched by all of its items.

A *generalized association rule* Srikant and Agrawal [1995] is an implication $A \rightarrow B$, where A and B are disjoint generalized itemsets, i.e., generalized itemsets having no attributes in common. Hereafter, A and B will be denoted as the antecedent and consequent of rule $A \rightarrow B$, respectively. Generalized rules are characterized by three main quality index, i.e., support, confidence, and lift.

The *support* of a rule $A \rightarrow B$ ($\mathrm{s}(A \rightarrow B)$) corresponds to the support of itemset $A \cup B$ in the analyzed dataset. It indicates the observed frequency of occurrence of the rule. High-support rules represent recurrent patterns that are likely to occur in the analyzed data not by chance.

The *confidence* of a rule $A \rightarrow B$ ($\mathrm{c}(A \rightarrow B)$) is the conditional probability of occurrence of generalized itemset B given generalized itemset A. It indicates the strength of the implication ($\rightarrow$) and it is computed as the percentage of records matched by all the items in A and B (i.e., the support of the rule) over the number of records matched by items of the rule antecedent (i.e., the support of A).

The *lift* of a generalized association rule $A \rightarrow B$ is defined as $lift(A,B) = \frac{c(A \rightarrow B)}{s(B)} = \frac{s(A \rightarrow B)}{s(A)s(B)}$ Tan et al. [2005], where $\mathrm{s}(A \rightarrow B)$ and $\mathrm{c}(A \rightarrow B)$ are respectively the rule support and confidence, and $\mathrm{s}(A)$ and $\mathrm{s}(B)$ are the supports of the rule antecedent and consequent. If lift(A,B)=1, the itemsets A and B are not correlated, i.e., they are statistically independent. Lift values below 1 show negative correlation, while values above 1 indicate a positive correlation between itemsets A and B.

The mining problem The GECkO system extracts from the prepared relational dataset all the generalized rules that satisfy a minimum support threshold *minsup* and a minimum confidence threshold *minconf*. Since both positively and negatively correlated rules are considered for in-depth analysis, no minimum/maximum lift threshold is enforced. While positively correlated rules represent strong correlations among data items, negatively correlated ones represent implications that hold less than expected.

The algorithms The generalized association rule mining task is accomplished as a two-step process: (i) Frequent generalized itemset mining, which extracts all the generalized itemsets whose support is above *minsup*. (ii) Generalized association rule mining, which extracts all the generalized rules whose support is above *minsup* and whose confidence is above *minconf*, starting from the previously mined set of frequent itemsets.

To accomplish Step (i), the GenIO algorithm is integrated in the GECkO system, while to perform Step (ii) the RuleGen procedure integrated in the Apriori algorithm is adopted. To prevent generating all the possible item combinations, GenIO generates a subset of potentially interesting generalized itemsets covering, at a higher abstraction level, most of the information represented by infrequent itemsets. More details on the GenIO and Apriori algorithms are given in Baralis et al. [2010] and Agrawal and Srikant [1994], respectively.

Rule categorization Exploring the results of the rule extraction process can be a challenging task, because the number of mined rules can be very high. To ease the manual exploration of the result, rules are categorized into a subset of classes according to the represented knowledge. Thus, experts can focus their attention on the subset of classes of interest.

Rule class Pollutant-Pollutant (PP). This class comprises all the rules that contain only items belonging to attributes related to pollutant concentration levels. Rules (PM_{10}, *red*) $\rightarrow$ ($PM_{2.5}$, *red*) and (PM_{10}, *yellow*) $\rightarrow$ (O_3, *non-critical*) are examples of rules of class PP. These rules can be useful for identifying correlations between the concentration levels of multiple pollutant and, thus, to plan targeted actions (e.g., planning air monitoring protocols, saving measurement costs).

Rule class Pollutant-Traffic (PT). This class comprises all the rules that contain items related to pollutant concentration levels and traffic conditions (e.g., number of gasoline engine vehicles). Rule (PM_{10}, *red*) $\rightarrow$ (*number of gasoline engine vehicles*, *high*) is an example of rules of class PT. These rules can be useful for correlating pollutant concentrations with the transit of different types of vehicles in the city. Based on these correlations, municipality managers may redesign traffic policies with the aim at reducing pollutant concentrations.

Rule class Pollutant-Meteo (PM). This class comprises all the rules that contain items related to pollutant concentration levels and meteorological conditions (e.g., temperature, humidity). Rule (PM_{10}, *red*) $\rightarrow$ (*temperature*, *very cold*) is an example of rules of class PM. These rules can be useful for correlating pollutants with climate conditions. Hence, they can identify meteorological conditions in which specific pollutants should be carefully monitored to prevent unsafe air conditions.

Rule class Pollutant-Date (PTE). This class comprises all the rules that contain items related to pollutant concentration levels and temporal attributes (e.g., date, time). Rule (PM_{10}, *red*) $\rightarrow$ (*date*, *morning*) is an example of rules of class PTE. These rules can be useful for correlating the levels of pollutants with specific time periods or time slots. Based on these rules, in-site monitoring actions can be scheduled at the timeslots at which high pollutant concentrations are most likely.

More complex rules, e.g., class Pollutant-Meteo-Traffic (PMT), can be extracted as well. They represent implications between pollutant levels and a combination of environmental conditions (e.g., rule (PM_{10}, *red*) $\rightarrow$ {(*temperature*, *very cold*), (*number of gasoline engine vehicles*, *high*)}).

Classes are manually explored by domain expert to infer potentially interesting knowledge from the contained rules. To consider first the top correlated combinations of pollutant data, rules are sorted by decreasing lift.

3 Experimental validation

The proposed approach was validated on real data acquired in Milan, that is one of the largest and most important Italian Smart Cities. To perform our analyses, we considered two open datasets collecting the sensor measurements acquired over a 12-month time period (i.e., over year 2013). The generalized rules were extracted by using the Python implementation of the GenIO algorithm Baralis et al. [2010] provided by the respective authors. The main algorithm characteristics are described in Section 2.3. We extracted frequent and high-confidence rules, which represent recurrent and potentially reliable correlations among multiple data items. Whenever not otherwise specified, the following standard parameter setting will be considered: *minsup*=1% and *minconf*=20%. The experiments were performed on a quad-core 3.30 GHz Intel Xeon workstation with 16 GB of RAM, running Ubuntu Linux 12.04 LTS.

3.1 Datasets

The analyzed datasets collect pollutant concentrations, climate conditions and traffic levels of different categories of vehicles acquired in the central area of Milan (zone C). The main dataset attributes and the taxonomy used to aggregate the data values at multiple abstraction levels are described in Section 2.2. The first dataset, hereafter denoted as *DailyMeasures*, collects the daily pollutant levels measured on a daily basis as well as the environmental information about meteorological and traffic conditions. The second dataset (*HourlyMeasures*) collects the hourly pollutants levels together with the corresponding environmental conditions.

Pollutant data were gathered by the ARPA Lombardia ARPA. Piedmont Region. through monitoring stations equipped with a set of sensors, each one measuring a different pollutant. Meteorological measurements were collected through the Weather Underground web service Wikipedia Meteo information about metereological data, which gathers data from a geo-referenced network of Personal Weather Stations (PWSs) registered by users. We considered three PWSs located in the city center. Traffic data were provided by the Municipality of Milan[1]. They consist of the counts of the number of vehicles entering in the central area of Milan, separately for each category of vehicles.

3.2 Knowledge discovery

The extracted rules were categorized, according to the type of item correlations they represent, into the classes described in Section 2.3. For each class, a subset of the most interesting rules extracted from both datasets is reported in Table 1. Some of the selected rules recall established correlations between pollutants and environmental factors, discussed by previous works on the topic (e.g. Lodovici et al. [2003], Elminir [2005], Statheropoulos et al. [1998]). However, as discussed below, the mined generalized association rules provide more insightful information than the ground knowledge, because they indicate the levels at which pollutants, climate factors, and traffic conditions are actually influenced with each other.

Correlations between pollutant levels (Class PP). When particulate matters PM_{10} and $PM_{2.5}$ have the same criticality level (e.g., *yellow*, *green*), a strongly positive pairwise item correlation appears (see Rules R_1-R_3). The positive rule lift values confirm that the pollutant levels co-occur more than expected. On the other hand, opposite pollutant levels (e.g., in Rule R_4 *green* for PM_{10} and *blue* for $PM_{2.5}$) show a negative correlation, meaning that the occurrence of a pollutant level implies the absence of the other one. Beyond pointing out the established correlation between the concentrations of particulate matters PM_{10} and $PM_{2.5}$, rules R_1-R_3 provide additional and potentially useful information, because they indicate the levels at which the two pollutants are most likely to be correlated with each other. The confidence of the aforesaid rules indicates the probability of occurrence of a pollutant level given the level of another pollutant. For example, according to the confidence value of Rule R_1, the probability of having level *yellow* for $PM_{2.5}$ given level *yellow* of PM_{10} is approximately 73%. These probabilities can be useful for planning air quality monitoring activities. For example, if two pollutants have a high probability of sharing levels *orange* and *red*, a critical concentration of one pollutant should trigger prompt monitoring actions targeted to the other pollutant as well.

Rules R_5 and R_6 show the correlation between PM_{10} and the pair Ozone (O_3) and carbon monoxide (CO). For example, a fairly critical level of PM_{10} (*yellow*) is often related to a *non-critical* level of Ozone. Rules R_5 and R_6 contain two generalized items each, i.e., (O_3, *non-critical*) and (CO, *highly critical*), which aggregate

[1] http://dati.comune.milano.it/

the information provided by their corresponding lower-level items (O_3, *blue*) and (*CO*, *orange*) at a higher abstraction level.

Rules R_7 and R_8 show the inverse relationship between the levels of Nitrogen dioxide (NO_2) and Ozone (O_3). The oxidation in atmosphere of Nitrogen oxide, Ozone, and other pollutants produces Nitrogen dioxide. Hence, a high concentration of Nitrogen dioxide is often associated with a low concentration of Ozone (and vice versa).

Correlations between pollutant levels and meteorological factors (Class PM). Rules R_9-R_{11} show a positive correlation between the external temperature values and the concentration of particulate matters PM_{10} and $PM_{2.5}$, Carbon Monoxide, and Ozone (O_3). For example, according to Rule R_{11}, when the temperature is cold and the precipitations are too weak to disperse the pollutants in the air, the concentrations of the aforesaid pollutants are likely to be fairly critical (i.e., levels *fairly high* for *CO* and *red* for $PM_{2.5}$, respectively). Conversely, for pollutant NO_2 an opposite trend comes out. In fact, based on generalized rule R_{15} (reported in the following for rule class PTE) the concentrations of NO_2 is low (level *blue*) in winter. On the other hand, when the temperature is hot or very_hot, the pollutant levels are likely to *non-critical* (i.e., *green* or *blue*).

Correlations between pollutant levels and time (Class PTE). Based on the lift value of Rules R_{12} and R_{13} (approximately one), pollutant levels and day of the week categories (i.e., weekday, weekend) seem to be statistically independent with each other. On the other hand, a correlation between pollutant levels and seasons holds. This effect seems to be an indirect consequence of the strong correlation holding between pollutant levels and temperature values.

Correlations between pollutant levels and traffic conditions (Class PTR). The effect of traffic flows on the air quality can be investigated by analyzing the rules involving pollutant levels and traffic conditions. For example, rules R_{20}-R_{22} show the correlation between the presence of many diesel engine vehicles in the city area and the concentration of PM_{10}. According to these rules, the presence of a medium/high number of vehicles is negatively correlated with a low concentration of PM_{10} and positively correlated with a fairly high concentration of the same pollutant. Conversely, a high number of gasoline engine vehicles is positively correlated with a low concentration of Carbon Monoxide. The latter rule indicates that the presence of diesel engine vehicles is critical for PM_{10} emissions, whereas gasoline engine vehicles does emit a significant amount of Carbon Monoxide.

4 Conclusions and future works

In this paper we presented a new data mining system to analyze air pollution-related data through generalized association rules. Preliminary results on real datasets demonstrate the potential of the proposed methodology in modeling interesting correlations at different abstraction levels. There is still room for improvements for our system. For example, GECKO may be enriched with (i) other kinds of interesting data affecting air quality such as people's mobility and private/public transport data, and (ii) data mining algorithms to discover correlations among weighted air pollution-related data.

5 Acknowledgment

The research leading to these results has received funding from the Italian Ministry of Research (MIUR) under the cluster *Tecnologie per le Smart Communities, Progetto MIE - Mobilità Intelligente Ecosostenibile.*

References

RapidMiner. Last access: February 2016. Online, 2016. URL http://rapid-i.com/content/view/181/190/.

R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th VLDB conference*, pages 487–499, 1994.

ARPA. Piedmont Region. *Regional Agency for the Protection of the Environment. Available at http://www.arpa.piemonte.it/english-version Last access: December 2014.*

Elena Baralis, Luca Cagliero, Tania Cerquitelli, Vincenzo D'Elia, and Paolo Garza. Support driven opportunistic aggregation for generalized itemset extraction. In *5th IEEE International Conference on Intelligent Systems, IS 2010, 7-9 July 2010, University of Westminster, London, UK*, pages 102–107, 2010.

Luca Cagliero and Paolo Garza. Improving classification models with taxonomy information. *Data Knowl. Eng.*, 86:85–101, 2013. doi: 10.1016/j.datak.2013.01.005. URL http://dx.doi.org/10.1016/j.datak.2013.01.005.

Luca Cagliero, Tania Cerquitelli, Paolo Garza, and Luigi Grimaudo. Twitter data analysis by means of strong flipping generalized itemsets. *Journal of Systems and Software*, 94:16–29, 2014.

Hamdy K. Elminir. Dependence of urban air pollutants on meteorology. *Science of The Total Environment*, 350(1-3):225 – 237, 2005. ISSN 0048-9697. doi: http://dx.doi.org/10.1016/j.scitotenv.2005.01.043.

M Lodovici, M Venturini, E Marini, D Grechi, and P Dolara. Polycyclic aromatic hydrocarbons air levels in florence, italy, and their correlation with other air pollutants. *Chemosphere*, 50(3):377–382, January 2003. ISSN 0045-6535. doi: 10.1016/s0045-6535(02)00404-6.

R. Srikant and R. Agrawal. Mining generalized association rules. In *VLDB 1995*, pages 407–419, 1995.

M Statheropoulos, N Vassiliadis, and A Pappa. Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment*, 32(6):1087 – 1095, 1998. ISSN 1352-2310.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Intoduction to Data Mining*. Addison Wesley, 2005.

Wikipedia Meteo information about metereological data. *Available at https://en.wikipedia.org/wiki/Rain, https://en.wikipedia.org/wiki/Wind, https://en.wikipedia.org/wiki/Ultravioletindex, https://en.wikipedia.org/wiki/Atmosphericpressure. Last access: February 2016.*

Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: when urban air quality inference meets big data. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1436–1444, 2013.

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2267–2276, 2015.

Figure 1: Rule examples.

Rule ID	Dataset	Rules	Sup (%)	Conf (%)	Lift
Class PP					
R_1	DailyMeasures	(PM_{10},*yellow*) → ($PM_{2.5}$,*yellow*)	9.7	72.9	5.4
R_2	DailyMeasures	(PM_{10},*green*) → ($PM_{2.5}$,*green*)	27.0	66.7	2.2
R_3	DailyMeasures	(PM_{10},*blue*) → ($PM_{2.5}$,*blue*)	37.78	95.78	2.0
R_4	DailyMeasures	(PM_{10},*green*) → ($PM_{2.5}$,*blue*)	10.3	25.2	0.52
R_5	DailyMeasures	(PM_{10},*yellow*) → {(O_3,*non-critical*), (*CO*,*highly critical*)}	7.0	52.1	5.1
R_6	DailyMeasures	(PM_{10},*yellow*) → {(O_3,*non-critical*), (*CO*,*highly critical*)}	5.6	41.7	5.0
R_7	HourlyMeasures	(O_3,*highly critical*) → (NO_2,*non-critical*)	5.9	51.6	1.5
R_8	HourlyMeasures	(O_3,*non-critical*) → (NO_2,*highly critical*)	11.3	24.1	1.7
Class PM					
R_9	DailyMeasures	{(*precipitations*,*drizzling*), (PM_{10},*orange*), ($PM_{2.5}$,*red*)} → {(*temperature*,*very cold*), (*CO*,*fairly high*)}	1.1	40.0	20.1
R_{10}	DailyMeasures	{(*temperature*,*very cold*), (*CO*,*fairly high*) → {(O_3,*blue*), ($PM_{2.5}$,*red*)}	1.4	55.6	2.2
R_{11}	DailyMeasures	{(*precipitations*,*no rain*), (*temperature*,*hot*), (C_6H_6,*non-critical*) } → ($PM_{2.5}$,*green*), (O_3,*non-critical*)}	1.1	66.7	20.7
Class PTE					
R_{12}	DailyMeasures	(PM_{10},*green*) → (*date*,*weekday*)	30.3	74.2	1.1
R_{13}	DailyMeasures	(PM_{10},*green*) → (*date*,*weekend*)	10.6	25.9	0.9
R_{14}	DailyMeasures	(*date*,*spring*) → (PM_{10},*green*)	5.6	55.6	1.4
R_{15}	DailyMeasures	(*date*,*winter*) → (NO_2,*blue*)	52.8	54.3	1.2
R_{16}	HourlyMeasures	(*hourly time period*,*late afternoon*) → (NO_2,*fairly critical*)	6.5	39.1	1.4
R_{17}	HourlyMeasures	(*hourly time period*,*night*) → (NO_2,*fairly critical*)	9.7	29.1	0.9
R_{18}	HourlyMeasures	(*hourly time period*,*late morning*) → (NO_2,*fairly critical*)	6.5	39.9	1.4
R_{19}	HourlyMeasures	{(*date*,*winter*), (O_3,*non-critical*)} → (NO_2,*highly critical*)	5.9	26.9	1.9
Class PTR					
R_{20}	DailyMeasures	(*num. diesel engine vehicles*,*medium*) → (PM_{10},*fairly high*)	9.7	70.0	1.8
R_{21}	DailyMeasures	(*num. diesel engine vehicles*,*high*) → (PM_{10},*green*)	14.7	34.4	0.9
R_{22}	DailyMeasures	(*num. gasoline engine vehicles*,*high*) → (CO,*low*)	9.2	73.3	1.4